LabGym: A versatile computational tool for automatic quantification of user-defined animal behaviors

Yujia Hu[1,*], Carrie R. Ferrario[2], Anjesh Ghimire[3], Brendon Watson[3], Kenichi Iwasaki[1], Hope White[1], Yitao Xi[1], Jie Zhou[4,*], Bing Ye[1,*]

[1] Life Sciences Institute and Department of Cell and Developmental Biology, University of Michigan, Ann Arbor, MI 48109, USA

[2] Department of Pharmacology and Psychology Department (Biopsychology), University of Michigan, Ann Arbor, MI 48109, USA

[3] Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109, USA

[4] Department of Computer Science, Northern Illinois University, DeKalb, IL 60115, USA

* Correspondence: Yujia Hu (henryhu@umich.edu), Jie Zhou (jzhou@niu.edu) and Bing Ye (bingye@umich.edu)

ABSTRACT

Quantifying animal behavior is important for many branches of biology and biomedical research. Current computational tools for this purpose typically rely on a limited number of pre-defined metrics to identify a behavior and then use these metrics as measurements of the behavior. However, categorizing behaviors by pre-defined metrics has limitations in its applicability (e.g., behavior types or animal species) and accuracy. Here we report a new tool, *LabGym*, for quantifying animal behaviors without such limitations. *LabGym* combines multi-animal tracking and deep learning algorithms with new designs to achieve high-throughput and accurate quantifications of each user-defined animal behavior. It also provides users a way to generate visualizable benchmark datasets for these behaviors, which are valuable resources for research in neuroscience, behavioral science, and computer science. We demonstrate the utility of *LabGym* in various user-defined behaviors in animal species ranging from soft-bodied invertebrates to mammals.

INTRODUCTION

Quantitative measurements of animal behavior are important for many branches of biology and biomedical research (Altimus et al., 2020; Datta et al., 2019; Egnor and Branson, 2016; Pereira et al., 2020). Current computational tools for behavioral quantification typically rely on a limited number of pre-defined metrics to identify a behavior (Dankert et al., 2009; Denisov et al., 2013; Hsu and Yttri, 2021; Kabra et al., 2013; Klibaite et al., 2017; Klibaite and Shaevitz, 2020; Mathis et al., 2018; Nilsson et al., 2020; Pereira et al., 2019; Segalin et al., 2020). These behavior-type-defining metrics are then used as measurements of the behavior. For example, JAABA (Kabra *et al.*, 2013) uses animals' instantaneous speed, positions, and areas for identifying a set of behaviors in mice and flies. Similarly, B-SOiD (Hsu and Yttri, 2021) utilizes mouse body part positions computed by DeepLabCut (Mathis *et al.*, 2018) to calculate speed, angles, and body lengths for categorizing behaviors specifically in rodents. In *Drosophila* larvae, SALAM/LARA evaluates specific features of larval contours to detect several pre-defined behaviors (Denisov *et al.*, 2013).

While tools that rely on pre-defined metrics have greatly advanced behavioral analyses, they have several limitations. First, the metrics need to be mathematically defined *a priori* to fit certain behavioral categories in specific animal species. However, animals can develop new behavior types or display different variance in the expression of a given behavior as a function of experimental manipulation, such as drug exposure (Ellinwood and Balster, 1974; Ferrario et al., 2005). The pre-defined metrics may fail to identify these new types of behavior. For example, moving speed that is used to distinguish running from walking may not be useful to define a behavior during which the animal is immobile, such as facial grooming. Moreover, the same behavior may be recorded from different perspectives, such as side view versus top view. The pre-defined metrics that are built for side view might not work for top view at all. Even worse, the metric-based approaches assume that only the metrics selected by the designer are key to identifying a given behavior. Nevertheless, the limited number of selected metrics typically underrepresent the complete information needed to capture the complexity of animal behaviors.

An alternative approach to quantification of animal behavior is to first categorize the behavioral types without using pre-defined metrics and then quantify various aspects of each type of behavior. Actually, behavioral categorization, but not quantification, can be performed by humans due to their perceptual abilities for pattern recognition (Reed, 1972). Similar to the human perceptual process, recent advancements in deep-learning based algorithms offer the

possibility of categorizing behaviors without using pre-defined metrics. For example, a recent study used deep convolutional networks to analyze raw frames in a video for behavioral categorizations (Bohnslav et al., 2021). However, this tool only classifies the behaviors but can neither provide quantitative measurements of the behavioral metrics nor be applied in videos with multiple animals.

In this paper, we report a new tool, *LabGym*, for quantifying user-defined animal behaviors without restrictions on behavior types or animal species. People who analyze behaviors in a video typically ignore the irrelevant background and only track the behaving animals. They then evaluate and integrate spatiotemporal information with the animal's overall movement pattern to categorize the behavior. Inspired by this human cognitive process, we designed *LabGym* to ignore irrelevant background, track multiple animals, and then categorize the behavioral types accurately without using pre-defined metrics. It then uses the information from the behavioral categories and the tracked animal foregrounds to calculate diverse metrics for quantitative measurements of different aspects of each user-defined behavior. *LabGym* achieves accurate categorization of behaviors by employing new deep-learning algorithms to holistically assess both the spatiotemporal information and the animal's overall movement pattern. While most current computational tools require demanding computational capabilities, behavioral analyses by *LabGym* are efficient and high-throughput, and can be carried out with a common laptop computer.

To perform categorization tasks, deep neural networks need to be trained on examples of ground truth (i.e., labeled data) that decides the categorization accuracy (LeCun et al., 2015; Roh et al., 2021; Schmidhuber, 2015). Since the manual labeling might vary among individuals, benchmark datasets that can be visualized are needed for benchmarking the labeling to ensure accuracy and reproducibility. In fact, shared datasets with consensus labeling are also essential for benchmarking different algorithms, which can then facilitate the development of computational tools for behavioral analyses. However, such visualizable benchmark datasets for behaviors are largely missing due to lack of tools to generate them. *LabGym* remedies this by providing users with a way to generate visualizable, standard behavioral datasets that can be shared across the research community.

We demonstrate that *LabGym* precisely captures even subtle changes in user-defined behaviors in animals ranging from soft-bodied invertebrates to mammals. We also used it to generate visualizable behavioral datasets in these species to serve as benchmarks for researchers in neuroscience, behavioral science, and computer science. These are "living" datasets that can be added to and modified across time. Finally, *LabGym* includes real-time

visualizations of behavioral categorizations and a graphical user interface (GUI) containing flexible options for experimenters to use without the need of writing computer code.

RESULTS

The *LabGym* pipeline

LabGym was designed to mimic human assessment of animal behaviors. The goals for the software are to efficiently and accurately identify all the user-defined behavioral types (i.e., "categories") for each animal in an experimental session, and then provide quantitative measurements of each category.

To accomplish this, *LabGym* first uses a new method to effectively remove backgrounds in each video frame and tracks each animal within the video (Figure 1A). At each frame of the video, a module termed the 'Categorizer' determines which behavior an animal performs during a time window that spans from the current frame back to a user-defined number of frames prior ('time window'). To achieve this, two types of data are extracted at each video frame for each animal: a) a background-free animation ('animation') that spans the specified time window, and b) the imprint of the positional changes of the animal's body over time in the animation ('pattern image'; Figure 1B and Supplementary Video 1). These two pieces of data are then analyzed by the Categorizer that comprises three submodules: the 'Animation Analyzer', the 'Pattern Recognizer', and the 'Decision Maker' (Figure 1C). The Animation Analyzer analyzes all spatiotemporal details in each animation, whereas the Pattern Recognizer evaluates the animal's overall motion pattern in the paired pattern image (see details in "*Biologically plausible design of the Categorizer for holistic assessments*"). The Decision Maker integrates the outputs from the Animation Analyzer and Pattern Recognizer to determine the behavior category within the specified time window.

Because it is based on deep neural networks, the Categorizer needs to be trained on examples of ground truth (i.e., labeled data) for each behavior in order to conduct accurate categorizations. To train the Categorizer, users only need to 'tell' the Categorizer the behavior names of the training examples (animations and their paired pattern images; Figure 1B). To do so, users can use *LabGym* to export unsorted examples, and manually sort each pair of example into folders created under the behavior name (Figure 1D). Users then input these folders into *LabGym* to generate a labeled dataset for training the Categorizer.

After the categorizations, the quantification module called the 'Quantifier' (Figure 1E) uses information from the behavioral categories and the animal foregrounds to compute various metrics for each behavior (e.g., duration, speed, etc.; Figure 1F). The output also includes temporal raster plots of behavioral events (and their probabilities) and annotated videos for real-time visualization of behavioral categorization.

Tracking multiple animals in various backgrounds for high-throughput and efficient analyses

Behavioral experiments often involve behavior-irrelevant background—such as the enclosures for holding the animals or changes in illumination during optogenetic manipulation—that renders analyses inaccurate and inefficient. To address these issues, we designed a module in *LabGym* that removes backgrounds but retains all the pixels representing the animals in each video frame (i.e., foreground).

We reasoned that backgrounds in experimental conditions are often static, and that the pixel value in each background location would be stable over time (Supplementary Figure 1A). Therefore, we developed a method to find the stable value of each pixel in each video frame and use these values to reconstruct the static background (Figure 2A and see **Methods**). The static background is then subtracted from each frame to obtain the foreground, which represents the animal (Figure 2B). Compared with two state-of-the-art background subtraction methods, Local SVD Binary Pattern (LSBP) (Guo et al., 2016) and GSoC (Bradski, 2000) (implemented in Google Summer of Code 2017), our stable-value detection method reconstructed much cleaner backgrounds (with less traces of the animals) on behavioral videos of multiple animals and in diverse experimental settings (Figure 2B and Supplementary Figure 1B).

The tracking of each animal over time is performed based on the assumption that the Euclidean distance between the centers of the same animal in consecutive frames are always the smallest among all Euclidean distances between each possible pair of animal centers. Each tracked animal is assigned by a unique identity (ID) that links to a matrix for storing all information about that animal over time.

Multiple animals in the same enclosure sometimes collide with each other, which can cause tracking errors. To make the analysis fully automatic, the current version of *LabGym* excludes animals from tracking if users choose not to allow animal body entanglements (Figure 2C, Supplementary Videos 2 and **Methods**). This ensures that the analysis is only performed on reliably tracked individual animals. If body entanglements are allowed, *LabGym* can also analyze social behaviors in which animals have body contact (Figure 2C, Supplementary Videos 3 and **Methods**). In this scenario, merged animals are analyzed as one.

Notably, animals are reliably tracked by *LabGym* even in videos with unstable or shifting illuminations (Supplementary Video 4) because the changes in background illumination (e.g., during optogenetic stimulation) can be reconstructed separately.

Taken together, the effective background subtraction and reliable animal tracking in *LabGym* make it applicable to a wide range of behavioral videos without the need for specialized recording equipment.

Biologically plausible design of the Categorizer for holistic assessments

We built the Categorizer with three submodules: Animation Analyzer, Pattern Recognizer, and Decision Maker (Figure 3 and **Methods**), each of which consists of neural networks that suit the design goals. Importantly, the two subnetworks—the Animation Analyzer and the Pattern Recognizer—were designed to analyze two types of behavioral information that complement each other, which ensures accurate categorizations of behavior through holistic assessment.

Animal behavior involve changes in both spatial and temporal dimensions. To learn the spatial details of the animal in each time point (frame) during a behavior, the Animation Analyzer uses time-distributed convolutional layers to analyze raw pixel values that represent the animal foreground in each frame of an animation (Figure 3A). To learn about how these frame-wise spatial details are organized along the temporal axis, the Animation Analyzer then uses recurrent layers to compute temporal connectivity among the frame-wise spatial details analyzed by time-distributed convolutional layers.

In the Animation Analyzer, maximum pooling operations were used to enable translational invariance (LeCun and Bengio, 1995), which eliminates the information of an animals' location in each frame so that the Animation Analyzer can be trained to recognize the animal regardless its location (e.g., left side vs. right side) in the frame. However, the information of the animal's locational changes during the behavior, which is key to behavioral categorizations, is also lost. To compensate this issue of the Animation Analyzer, we generated a "pattern image" for each animation to depict the animal's overall movement and designed the Pattern Recognizer (that comprises convolutional layers) to analyze the pattern image (Figure 3B). In each pattern image, the contours of both the whole body and the body parts of the animal in each frame of the animation were superimposed with gradually changed colors to indicate their temporal sequences. This shows the animal's overall movement pattern during the animation in a single 2-dimensional image to allow for efficient analyses.

Finally, the Decision Maker uses concatenating layers to integrate the outputs from both the Animation Analyzer and the Pattern Recognizer, and passes the integrated information to fully-connected (dense) layers for discriminating behavioral categories (Figure 3C).

Users can choose to include background in the animations as it sometimes can be useful for behavioral categorization. For example, bedding in a mouse's mouth can indicate nest

building and, thus, including the bedding in the animations can facilitate the categorization of nest building behavior (Figure 4A). Users can also choose whether to show the animals' body parts in the pattern images. Whole-body motion patterns are sometimes sufficient for distinguishing behaviors like curling versus uncoiling in *Drosophila* larvae, wing extension versus abdomen bending in adult flies, or walking versus rearing in rats (Figure 4B), while body part motion patterns are key for distinguishing similar behaviors like facial grooming versus sniffing, or ear grooming versus sitting in mice (Figure 4C).

The Categorizer accurately categorizes user-defined behaviors across species

Since the behavior types for the Categorizer to recognize are defined by users, rather than pre-defined in the program, the Categorizer can identify behaviors without restrictions on behavioral categories or animal species. To demonstrate this, we used *LabGym* to study relatively simple behaviors in *Drosophila* larvae and behaviors of increasing complexity in rodents. We first used *LabGym* to generate and label separate behavioral datasets for *Drosophila* larvae (Supplementary Video 5), rats (Supplementary Video 6), and mice (Supplementary Video 7). These datasets are publicly available (see **Methods**) and can evolve as anyone who works on animal behaviors can check the labeling and add their data to expand and diversify them. The larva dataset contains the fewest categories with sufficient examples to cover most behavioral variants; the rat dataset contains more categories of more complicated behaviors that only show subtle differences among each other; the mouse dataset contains even more categories of even more complex behaviors. To test the robustness of *LabGym*, we also varied the amount of training examples used to categorize a given behavior, ranging from ~100 to ~800 pairs of animations and pattern images.

The Categorizer is customizable with different levels of complexity to address behaviors of various complexity. Therefore, users can set and train the Categorizer for their own behavioral datasets. To provide a reference for customizing the Categorizer, which is a trial-and-error process, we trained several Categorizers with different complexity for each species, and determined the most suitable ones based on their overall accuracy of behavioral categorizations in training (Supplementary Table 1 and **Methods**). The selected Categorizer for each dataset were named as *LarvaNociceptor_Topview_30fps* (*LarvaN*), *MouseHomecage_Sideview_30fps* (*MouseH*) and *RatAddiction_Topview_30fps* (*RatA*). To test the accuracy and generalizability of these three Categorizers, we assessed their precisions, recalls (sensitivity), and F1 score (weighted precision and sensitivity, see **Methods**) of

behavioral categorizations in videos that were not used for generating the training or validation dataset.

Complex deep neural networks such as ResNet50 (He et al., 2016), which have been used in previous behavioral analysis tools (Bohnslav *et al.*, 2021; Mathis *et al.*, 2018), might not be suitable for simple larval behaviors since these complex networks would learn too many irrelevant details in simple datasets, resulting in reduced accuracy (i.e., overfitting). In contrast, *LarvaN* with simple complexity (Supplementary Table 1) achieved 97% accuracy in frame-wise categorizations of 6 different behaviors elicited by noxious stimuli (Figure 5A and Supplementary Video 8).

Rats show complex behaviors and the differences among them can be extremely subtle due to the recording or experimental conditions (Supplementary Video 6). Despite these challenges, the *RatA* performed reliably well in frame-wise differentiation of visually similar behaviors such as body grooming versus face grooming, or head swaying versus in place (Figure 5B and Supplementary Video 9).

Animals can develop new behaviors when experimental conditions change. For example, mice in a cage equipped with a wheel tend to run on the wheel. The movement of running on the wheel has features and dynamics that are distinct from running across a flat surface. Thus, wheel-running might not be correctly identified by tools in which running is a predefined. Additionally, a camera placed at the side of a cage will result in at least two viewpoints of the mouse at it moves in the cage, a side view and a front view, which can greatly complicate the automated analysis of behavior. Despite these challenges, *MouseH* achieved excellent accuracy in frame-wise categorization of 20 behavioral categories with as few as 79 training examples for some categories (Figure 5C and Supplementary Video 10). Although both were used for categorizing rodent behaviors, *MouseH* trained on pattern images including the body parts (Supplementary Videos 7) performed better than *RatA* trained on those excluding body parts (Supplementary Videos 6; Figure 5B and 5C). This indicates that including the body parts in pattern images can boost the learning efficiency of the Categorizer (Figure 4 and **Methods**).

Due to the data-driven nature of deep learning, the accuracy and generalizability of the Categorizer largely depend on the amount, diversity, and labeling accuracy of the training examples. Therefore, the performance of the Categorizers can be continuously improved over time with the accumulation of well-labeled training examples.

The Quantifier provides quantitative measurements that are sufficient for capturing subtle changes in diverse aspects of each behavior

The quantitative measurements of behaviors in existing (metric-based) tools typically rely on the animal tracking data itself, making it difficult to obtain quantitative measurements that are specific to a certain behavior category. In contrast, since the behavioral categorization in *LabGym* is independent of the tracking data, the Quantifier in *LabGym* can use both the behavioral category output from the Categorizer and the spatial information of the tracked animals to calculate a wide range of measurements that are specific to each behavioral category. The current version of *LabGym* includes 14 measurements: acceleration/velocity reduction, angle, count, distance, duration, intensity (area), intensity (length), latency, magnitude (area), magnitude (length), speed, velocity, vigor (area) and vigor (length) (Figure 6 and **Methods**). Additional measurements can be added on demand.

To test how well the Quantifier performs in quantifying behaviors in diverse animal species, we applied *LabGym* to experiments in larvae, rats, and mice, in which the quantitative measurements of dynamics or intensity of the behaviors was difficult to achieve with previous methods.

Previous studies show that optogenetic inhibition of larval leucokinin (LK) neurons slightly impaired larval behaviors elicited by nociceptor activation (Hu et al., 2020). In these experiments, both the subtle changes in behaviors and the shifting illumination caused by optogenetic manipulations posed challenges for automatic behavioral analysis. We thus tested whether *LabGym* was able to capture the behavioral changes in these experiments (Supplementary Video 11). Compared with the control group, larvae in the LK inhibition group showed a modest reduction in the probability of body rolling behavior, which is a characteristic nociceptive response (Hwang et al., 2007; Tracey et al., 2003) (Figure 7A and 7B). No change was observed in the latency of the response onset after optogenetic stimulation of nociceptors (Figure 7C). Moreover, LK inhibition caused a slight yet significant reduction in rolling velocity (Figure 7D) and duration (Figure 7E). We also compared several quantitative measurements of different behaviors and found that the results indeed reflect the ethological meaning of these behaviors. For example, rolling achieves higher speed than crawling (Figure 7F), supporting the notion that rolling was a more efficient way than crawling to escape from harm (Hwang *et al.*, 2007; Tracey *et al.*, 2003). Similarly, despite their spatial similarity, curling was more vigorous than turning (Figure 7G), echoing the fact that larvae typically perform curling under noxious stimuli but tend to turn in in response to milder situation like gentle touch or exploration (Hwang *et al.*, 2007; Jovanic et al., 2016; Lahiri et al., 2011; Tracey *et al.*, 2003). Thus, *LabGym* is

sufficiently sensitive for detecting modest effects of in vivo neuronal inhibition on behavior, and it does so on videos with shifting illumination.

When *Drosophila* larvae are exposed to sound, they display a startle response that is characterized by the cessation of crawling ("freezing"), retraction of their heads ("hunching"), and excessive turning movements (Zhang et al., 2013). To test whether *LabGym* could identify the subtle differences in behaviors elicited by different intensities of sound, we trained another Categorizer (*LarvaMechanosensor_TopView_30fps*) with examples of sound-elicited behaviors in *Drosophila* larvae (Supplementary Videos 5) and used it for analysis (Supplementary Video 12). This Categorizer has the same complexity as *LarvaN*. Compared to larvae exposed to 80 decibel (dB) white noise, those exposed to 94 dB showed a similar hunching probability (Figure 7H and 7I), but the magnitude of head retraction during hunching was enhanced (Figure 7I). Furthermore, larvae exposed to 94 dB showed more frequent turning (Figure 7K) and prolonged freezing (i.e., immobility; Figure 7I), which resulted in significant reductions in crawling distance (Figure 7M). However, when larvae froze, the reduction in the velocity of their movement was similar between the two groups (Figure 7N). Thus, *LabGym* was able to capture quantitative differences in the behaviors induced by auditory stimuli of different intensities.

It is well-established that drugs like amphetamine, produce dose-dependent increases in "psychomotor activity", and that the intensity of these behaviors increases (i.e., sensitizes) with repeated drug exposure (Ellinwood and Balster, 1974; Ferrario *et al.*, 2005). Psychomotor activity in rats encompasses a wide range of behaviors that include rearing, hyper-locomotion, and stereotyped, repetitive head movements (i.e., stereotypy). Psychomotor activity is most often quantified using non-parametric rating scales based on clusters of behavioral features and their intensities (Ellinwood and Balster, 1974; Ferrario *et al.*, 2005). To date, no automated systems have been able to sufficiently categorize or quantify psychomotor activity beyond locomotion or rearing (Ellinwood and Balster, 1974; Ferrario *et al.*, 2005).

LabGym automatically captured dose-dependent changes in key behavioral features of psychomotor activity (Figure 8A and Supplementary Video 13). These include rearing (Figure 8B), locomotion (Figure 8C), and head movements (persistent head movements indicating stereotypy; Figure 8D), following systemic amphetamine administration. Impressively, *LabGym* even captured stereotypy interspersed with small bursts of locomotion—an indication of a strong psychomotor response at high dose of amphetamine (e.g., Figure 8A and 8C, 5.6 mg/kg dose group in amphetamine baseline group). *LabGym* also captured quantitative changes in these behaviors during sensitization, in which repeated injection of amphetamine resulted in more intense behavioral responses to the same dose of drug (Figure 8E-8H). In contrast, the intensity

of spontaneously occurring behaviors such as body/face grooming were stable across time in rats treated with repeated saline (Figure 8I-8K).

When an animal ages, its ethogram might not change, yet the quantitative measurements of its behaviors (e.g., running speed) may decline. The speed of a mouse running on a wheel can be measured only by tracking the rotation of the wheel itself, which poses a challenge for computational tools designed to only track animals. Nevertheless, *LabGym* not only categorized the complete behaviors for both younger (32-week) and older mice (80-week) in a cage with a running wheel (Figure 8L and Supplementary Video 14), but also captured the decline in movement speed of the older versus the young mice by a parameter (vigor_area) that summarizes the areal changes per unit time of the animal's body (Figure 8M, 6 and **Methods**).

Taken together, these results demonstrate that the Quantifier reliably and precisely captures quantitative changes in behaviors across species, and in a wide range of settings and experimental conditions.

GUI with optional settings for user-friendly usage

To make *LabGym* applicable in various experimental settings and accessible to users without requiring a software programming background, we implemented a GUI. This includes 4 functional units: Generate Datasets, Train Networks, Test Networks and Analyze Behaviors.

The Generate Datasets unit (Supplementary Figure 2) is for batch processing of videos to generate visualizable behavioral data (animations and paired pattern images) that can be used to establish benchmark datasets for training the neural networks. The Train Networks unit (Supplementary Figure 3) is for preparing training datasets and training Categorizers at various levels of complexity. The Test Networks unit (Supplementary Figure 4) is for testing the accuracy and generalizability of a trained Categorizer in an unbiased manner. The Analyze Behaviors unit (Supplementary Figure 5) is for batch-processing of videos for behavioral analysis. It generates an annotated video (Supplementary Videos 8-14), a raster plot (Figure 7A, 7H, 8A and 8L) of behavioral events (displayed in user-defined colors) and their occurring probabilities (by color intensities) for each animal in one batch of analysis, and an Excel/CSV file containing individual behavioral parameters for each behavior of each animal.

Notably, the flexible options in *LabGym*, especially the customizable neural network complexity, allows it to achieve fast processing speed in training networks and in analyzing behaviors on commonly used laptops that do not have GPUs (Supplementary Table 2). This is

difficult to achieve for tools using fixed, complex neural networks (Bohnslav *et al.*, 2021; Mathis *et al.*, 2018).

DISCUSSION

LabGym is a unique computational tool that enables a generalizable solution to behavioral analysis. The flexibility and trainability of this tool allows it to be applied to very broad variety of species, behaviors and questions, and to be iteratively refined towards a particular set of conditions over time. This tool will open opportunities for automatic, high-throughput, targeted and customizable analyses of behaviors without the cost of commercial products, which will greatly facilitate the process of mechanistic discoveries in biological research. The standardized, visualizable behavioral datasets generated by *LabGym* will also benefit research in neuroscience, behavioral science, and computer science by helping to standardize procedures within and across research groups, and by providing long-term references for future studies to build on.

Holistic assessments minimize information loss and enhance generalizability

While tools that rely on pre-defined metrics are valuable, and have greatly advanced behavioral analyses in many fields, embedded in such approaches is an assumption that the unused or lost information is not important for the behaviors. In contrast, *LabGym* uses a deep neural network with novel design for holistic assessments of complete spatiotemporal information based on all aspects of an animal's behavior, including whole-body movement and body part movement at once. This prevents the huge information loss in the process of summarizing these features, and also enables users to define and refine the behavioral features that are most appropriate to their studies.

In metric-based behavioral analysis tools, specific metrics need to be mathematically defined *a priori* to fit certain behaviors and might not be useful for other behaviors. For example, the speed of forward movement can be useful for categorizing larval crawling, but is useless for categorizing larval turning that is defined by the bending of the body axis. In fact, behaviors are often intrinsically different across different categories, and even more so across different species. In addition, under certain experimental conditions, animals can develop new behaviors, which might not be analyzable with the pre-defined metrics. In contrast, *LabGym* does not rely on pre-defined metrics and thus works for a diversity of user-defined behaviors in different species. Importantly, the behavioral categorization by *LabGym* is end-to-end; that is, users directly tell *LabGym* the behavioral categories by providing examples, and *LabGym* outputs the behavioral categories for verification and refinement. It does not require users to select metrics and integrate them into an algorithm as criteria for recognizing behavioral types; nor does it

require them to be experts in computation or programming to mathematically describe the metrics. Instead, users only need to recognize and sort video examples of a given behavior.

Recently, DeepEthogram has been developed to use deep neural networks to assess raw pixels for behavioral categorizations (Bohnslav *et al.*, 2021). However, this tool only categorizes behaviors but does not provide quantitative measurements such as the movement speed or magnitude of the behavior, which are key to behavioral analysis. Moreover, it is not suitable for analyzing videos that include multiple animals since it assesses every pixel in a video frame and thus cannot track animals. Moreover, computation on all pixels of an entire video frame is ineffective and inefficient. In fact, the noise in the optical flows caused by irrelevant background can be clearly visualized in DeepEthogram (Bohnslav *et al.*, 2021). In contrast, *LabGym* ignores the irrelevant background and can analyze multiple animals simultaneously, which is efficient and high throughput. It also provides several options of neural network complexity, which makes the analyses flexible to behaviors of different complexity. Importantly, *LabGym* provides diverse quantitative measurements for each user-defined behavior.

Generate benchmark behavioral datasets for the research community

In behavioral categorization based on deep learning approaches, the accuracy depends on the quality of training datasets, which should be diverse and well labeled (LeCun *et al.*, 2015; Roh *et al.*, 2021; Schmidhuber, 2015). In the computer vision field, high-quality datasets, such as ImageNet (Deng et al., 2009), have been established to facilitate the training of deep neural networks and benchmarking different algorithms. In neuroscience, however, such benchmark datasets for behaviors of model organisms are largely absent. A cause of this is the lack of tools that can help researchers to generate standardized, visualizable behavioral datasets that are easy to label and to share. Although transfer learning that applies acquired network weights to new categorization problems can potentially be used to compensate for the lack of species-specific behavioral datasets, its efficiency relies on the similarity between the existing datasets used for pretraining the networks and the new data (Yosinski et al., 2014). Current tools using transfer learning for behavioral categorizations (Bohnslav *et al.*, 2021) utilized network weights that were acquired from human behavioral datasets such as Kinetics-700 (Carreira et al., 2019). However, human behaviors are intrinsically different from the behaviors of model organisms, especially the soft-bodied ones. Therefore, species-specific behavioral datasets are still in critical demand for accurate behavioral categorizations in model organisms.

LabGym addresses the lack of such benchmark datasets by providing a way for users to generate standardized, visualizable behavioral datasets that are easy to label or to benchmark the labeling, which makes it convenient for different experts to check, discuss and reach consensus on the labeling. Importantly, such expert-labeled behavioral datasets, if shared in the research community, not only ensure the accurate and reproducible executions of categorization criteria across different experimenters or trials, but also provide benchmarks for comparing different algorithms or references for future studies to build on.

In summary, *LabGym* is a powerful and versatile computational tool for automatically categorizing and quantifying animal behaviors without restrictions on behavior types or animal species.

ACKNOWLEDGMENTS

We thank Drs. Daniel Leventhal and Guoqiang Yu for helpful comments on earlier versions of the manuscript. This research was supported by National Institutes of Health (NIH) to B.Y. (R01NS104299 and R01EB028159). *Drosophila* Stocks from the Bloomington *Drosophila* Stock Center (NIH P40OD018537) were used in this study. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

AUTHOR CONTRIBUTIONS

Y.H. and B.Y. conceived the project. Y.H. designed and programmed *LabGym* with the guidance of J.Z. and B.Y.. Y.H., K.I., H.W., C.R.F., A.G., and B.W. performed the behavioral experiments on nociceptor-stimulation-elicited behaviors (Y.H.) and sound-elicited behaviors (K.I. and H.W.) in *Drosophila* larvae, rats (C.R.F.), and mice (A.G. and B.W.). Y.H., K.I., H.W., and Y.X. generated and labelled the dataset for larval nociceptor-stimulation-elicited behaviors (Y.H. and Y.X.), larval sound-elicited behaviors (K.I. and H.W.), rat behaviors (Y.H.), and mouse behaviors (Y.H. and A.G.). Y.H. performed the computational analyses and the validation of behavioral categorizations with the assistance from C.R.F., A.G., B.W., K.I., and H.W.. Y.H., C.R.F., B.W., J.Z., and B.Y. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Altimus, C.M., Marlin, B.J., Charalambakis, N.E., Colon-Rodriquez, A., Glover, E.J., Izbicki, P., Johnson, A., Lourenco, M.V., Makinson, R.A., McQuail, J., et al. (2020). The Next 50 Years of Neuroscience. J Neurosci *40*, 101-106. 10.1523/Jneurosci.0744-19.2019.

Bohnslav, J.P., Wimalasena, N.K., Clausing, K.J., Dai, Y.Y., Yarmolinsky, D.A., Cruz, T., Kashlan, A.D., Chiappe, M.E., Orefice, L.L., Woolf, C.J., and Harvey, C.D. (2021). DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels. Elife *10*. ARTN e63377 10.7554/eLife.63377.

Bradski, G. (2000). The OpenCV library. Dr Dobbs J *25*, 120-+.

Carreira, J., Noland, E., Hillier, C., and Zisserman, A. (2019). A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987.

Dankert, H., Wang, L.M., Hoopfer, E.D., Anderson, D.J., and Perona, P. (2009). Automated monitoring and analysis of social behavior in Drosophila. Nat Methods *6*, 297-303. 10.1038/Nmeth.1310.

Datta, S.R., Anderson, D.J., Branson, K., Perona, P., and Leifer, A. (2019). Computational Neuroethology: A Call to Action. Neuron *104*, 11-24. 10.1016/j.neuron.2019.09.038.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Li, F.F. (2009). ImageNet: A Large-Scale Hierarchical Image Database. Cvpr: 2009 Ieee Conference on Computer Vision and Pattern Recognition, Vols 1-4, 248-255. DOI 10.1109/cvpr.2009.5206848.

Denisov, G., Ohyama, T., Jovanic, T., and Zlatic, M. (2013). Model-based Detection and Analysis of Animal Behaviors using Signals Extracted by Automated Tracking. pp. 175-181.

Egnor, S.E., and Branson, K. (2016). Computational Analysis of Behavior. Annu Rev Neurosci *39*, 217-236. 10.1146/annurev-neuro-070815-013845.

Ellinwood, E.H., Jr., and Balster, R.L. (1974). Rating the behavioral effects of amphetamine. Eur J Pharmacol *28*, 35-41. 10.1016/0014-2999(74)90109-5.

Ferrario, C.R., Gorny, G., Crombag, H.S., Li, Y.L., Kolb, B., and Robinson, T.E. (2005). Neural and behavioral plasticity associated with the transition from controlled to escalated cocaine use. Biol Psychiat *58*, 751-759. 10.1016/j.biopsych.2005.04.046.

Guo, L.L., Xu, D., and Qiang, Z.P. (2016). Background Subtraction using Local SVD Binary Pattern. Ieee Comput Soc Conf, 1159-1167. 10.1109/Cvprw.2016.148.

He, K.M., Zhang, X.Y., Ren, S.Q., and Sun, J. (2016). Deep Residual Learning for Image Recognition. Proc Cvpr Ieee, 770-778. 10.1109/Cvpr.2016.90.

Hsu, A.I., and Yttri, E.A. (2021). B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors. Nature Communications *12*, 5188. 10.1038/s41467-021-25420-x.

Hu, Y.J., Wang, C.C., Yang, L.M., Pan, G., Liu, H., Yu, G.Q., and Ye, B. (2020). A Neural Basis for Categorizing Sensory Stimuli to Enhance Decision Accuracy. Curr Biol *30*, 4896-+. 10.1016/j.cub.2020.09.045.

Hwang, R.Y., Zhong, L.X., Xu, Y.F., Johnson, T., Zhang, F., Deisseroth, K., and Tracey, W.D. (2007). Nociceptive neurons protect Drosophila larvae from parasitoid wasps. Curr Biol *17*, 2105-2116. 10.1016/j.cub.2007.11.029.

Jovanic, T., Schneider-Mizell, C.M., Shao, M., Masson, J.B., Denisov, G., Fetter, R.D., Mensh, B.D., Truman, J.W., Cardona, A., and Zlatic, M. (2016). Competitive Disinhibition Mediates Behavioral Choice and Sequences in Drosophila. Cell *167*, 858-+. 10.1016/j.cell.2016.09.009.

Kabra, M., Robie, A.A., Rivera-Alba, M., Branson, S., and Branson, K. (2013). JAABA: interactive machine learning for automatic annotation of animal behavior. Nat Methods *10*, 64-U87. 10.1038/Nmeth.2281.

Klibaite, U., Berman, G.J., Cande, J., Stern, D.L., and Shaevitz, J.W. (2017). An unsupervised method for quantifying the behavior of paired animals. Phys Biol *14*. ARTN 015006
10.1088/1478-3975/aa5c50.

Klibaite, U., and Shaevitz, J.W. (2020). Paired fruit flies synchronize behavior: Uncovering social interactions in Drosophila melanogaster. PLoS Comput Biol *16*, e1008230. 10.1371/journal.pcbi.1008230.

Lahiri, S., Shen, K.L., Klein, M., Tang, A.J., Kane, E., Gershow, M., Garrity, P., and Samuel, A.D.T. (2011). Two Alternating Motor Programs Drive Navigation in Drosophila Larva. Plos One *6*. ARTN e23180 10.1371/journal.pone.0023180.

LeCun, Y., and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks *3361*, 1995.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436-444. 10.1038/nature14539.

Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., and Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience *21*, 1281-+. 10.1038/s41593-018-0209-y.

Nilsson, S.R., Goodwin, N.L., Choong, J.J., Hwang, S., Wright, H.R., Norville, Z.C., Tong, X., Lin, D., Bentzley, B.S., and Eshel, N. (2020). Simple Behavioral Analysis (SimBA)–an open source toolkit for computer classification of complex social behaviors in experimental animals. BioRxiv.

Pereira, T.D., Aldarondo, D.E., Willmore, L., Kislin, M., Wang, S.S.H., Murthy, M., and Shaevitz, J.W. (2019). Fast animal pose estimation using deep neural networks. Nat Methods *16*, 117-+. 10.1038/s41592-018-0234-5.

Pereira, T.D., Shaevitz, J.W., and Murthy, M. (2020). Quantifying behavior to understand the brain. Nat Neurosci *23*, 1537-1549. 10.1038/s41593-020-00734-z.

Reed, S.K. (1972). Pattern Recognition and Categorization. Cognitive Psychol *3*, 382-407. Doi 10.1016/0010-0285(72)90014-X.

Roh, Y., Heo, G., and Whang, S.E. (2021). A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective. Ieee T Knowl Data En *33*, 1328-1347. 10.1109/Tkde.2019.2946162.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks *61*, 85-117. 10.1016/j.neunet.2014.09.003.

Segalin, C., Williams, J., Karigo, T., Hui, M., Zelikowsky, M., Sun, J.J., Perona, P., Anderson, D.J., and Kennedy, A. (2020). The Mouse Action Recognition System (MARS): a software pipeline for automated analysis of social behaviors in mice. BioRxiv.

Tracey, W.D., Wilson, R.I., Laurent, G., and Benzer, S. (2003). painless, a Drosophila gene essential for nociception. Cell *113*, 261-273. Doi 10.1016/S0092-8674(03)00272-1.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks ? Adv Neur In *27*.

Zhang, W., Yan, Z.Q., Jan, L.Y., and Jan, Y.N. (2013). Sound response mediated by the TRP channels NOMPC, NANCHUNG, and INACTIVE in chordotonal organs of Drosophila larvae. P Natl Acad Sci USA *110*, 13612-13617. 10.1073/pnas.1312477110.

Figure 1

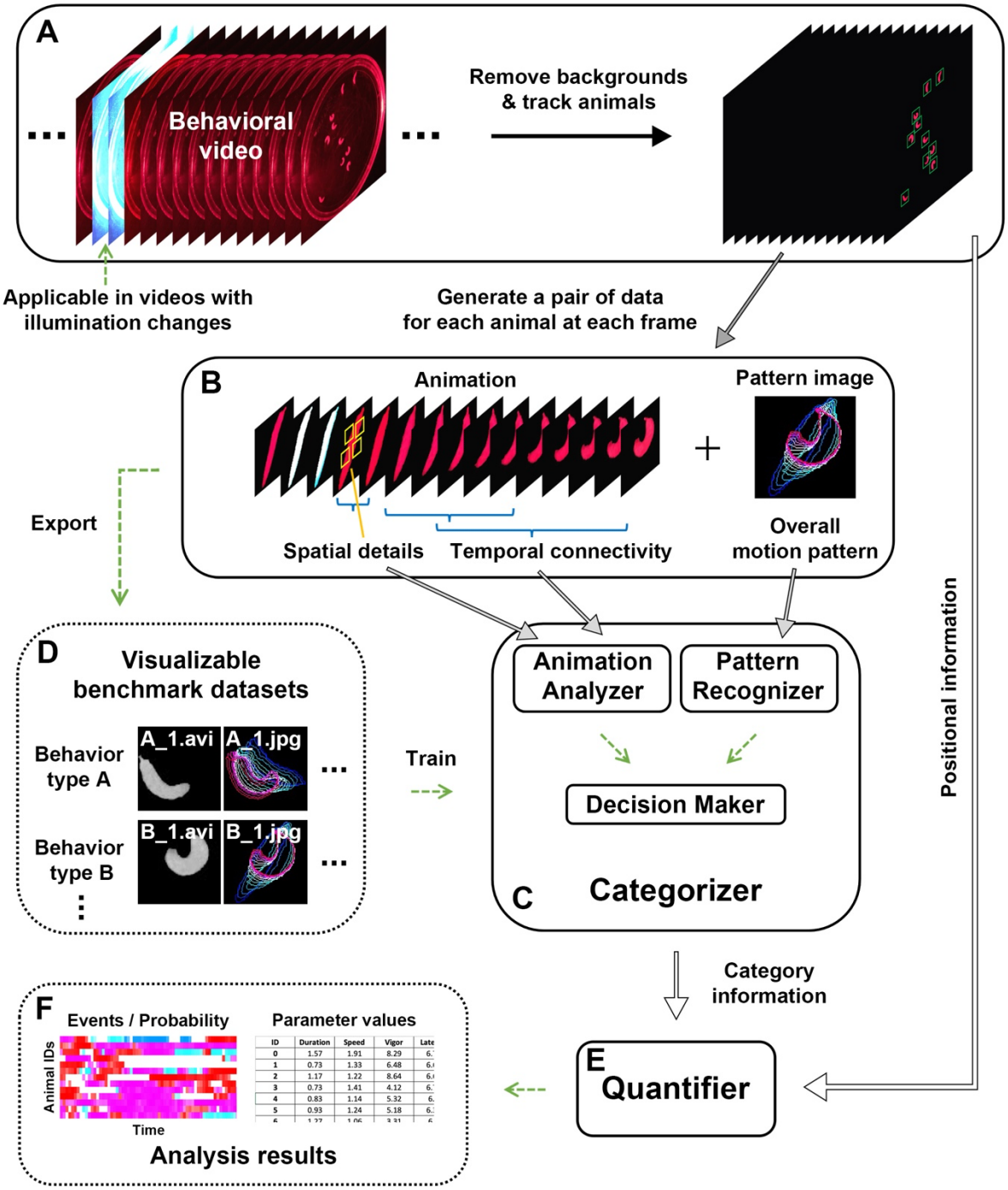


Figure 1. The pipeline of *LabGym*.

A. Irrelevant backgrounds are removed in each video frame and individual animals are tracked.

B. At each frame of the video, *LabGym* generates a pair of data for each animal, which comprises a short animation spanning a user-defined time window and its paired pattern image covering the overall pattern of the animal within the animation.

C. The Categorizer uses both the animations and pattern images to categorize behaviors during the user-defined time window (i.e., the duration of the animation). The Categorizer comprises three submodules, the ‘Animation Analyzer’, the ‘Pattern Recognizer’, and the ‘Decision Maker’. The Animation Analyzer analyzes all spatiotemporal details in the animation for each animal whereas the Pattern Recognizer evaluates the overall movement patterns in the paired pattern image. The Decision Maker integrates the outputs from both the Animation Analyzer and the Pattern Recognizer to determine which behavioral category is present in the animation for each animal.

D. The animations and pattern images can be exported to build benchmark datasets for training the Categorizer.

E and **F**. After the behavioral categorizations, the quantification module (‘Quantifier’) uses information from the behavioral categories and the animal foregrounds to compute specific quantitative measurements for different aspects of each behavior.

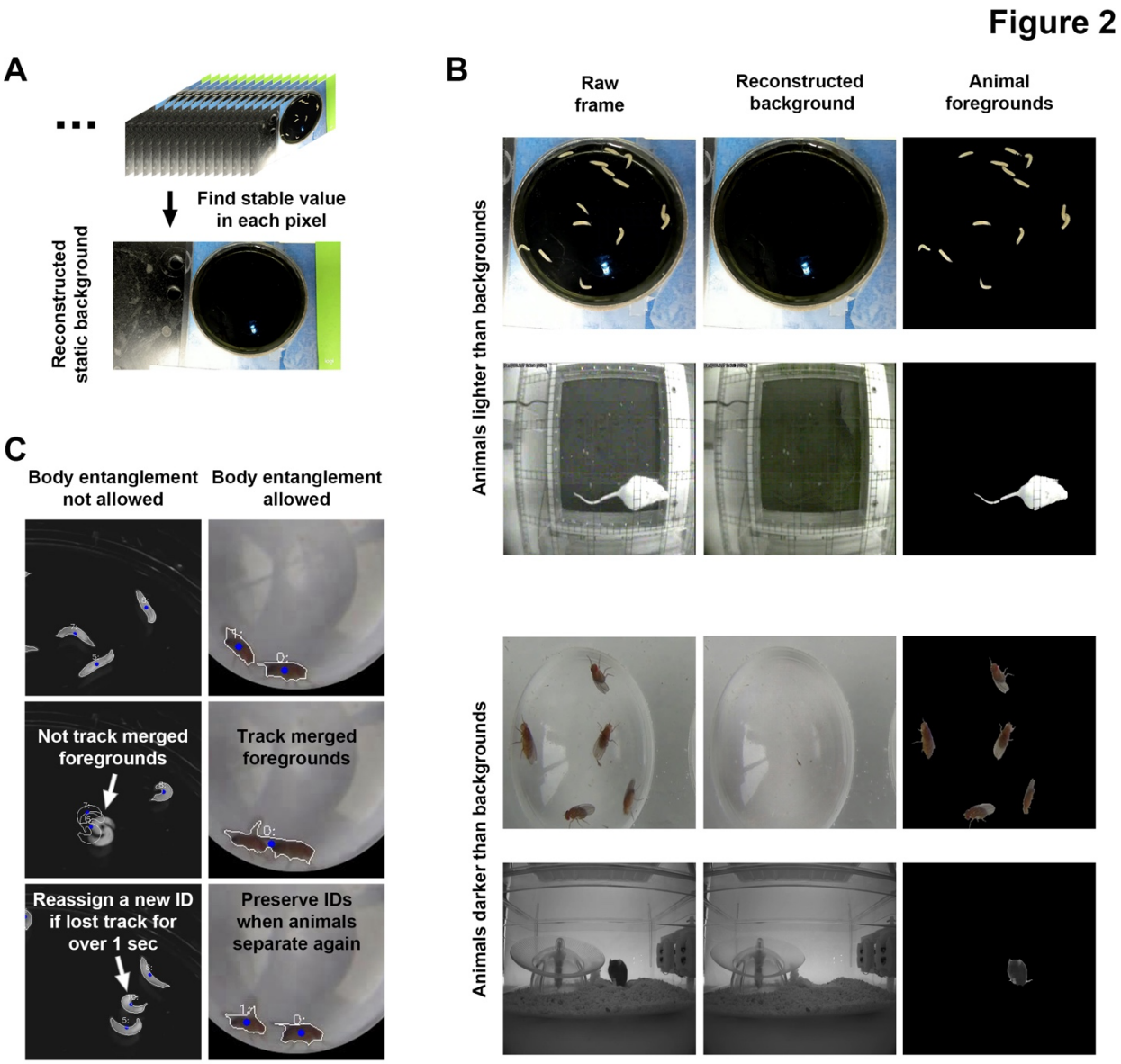


Figure 2. *LabGym* removes backgrounds and accurately tracks multiple animals.

A. The stable value of each pixel is used to reconstruct the static background for a video.

B. Examples of video frames showing that animal foregrounds are clearly presented after removing the static backgrounds that are constructed by our method of stable-value detection.

C. Examples of video frames showing that *LabGym* accurately track multiple animals performing either non-interactive or social behavior. Left: non-interactive behavior, in which animal (larva) contact is not allowed; right: social interaction in which two animals (adult flies) contact each other.

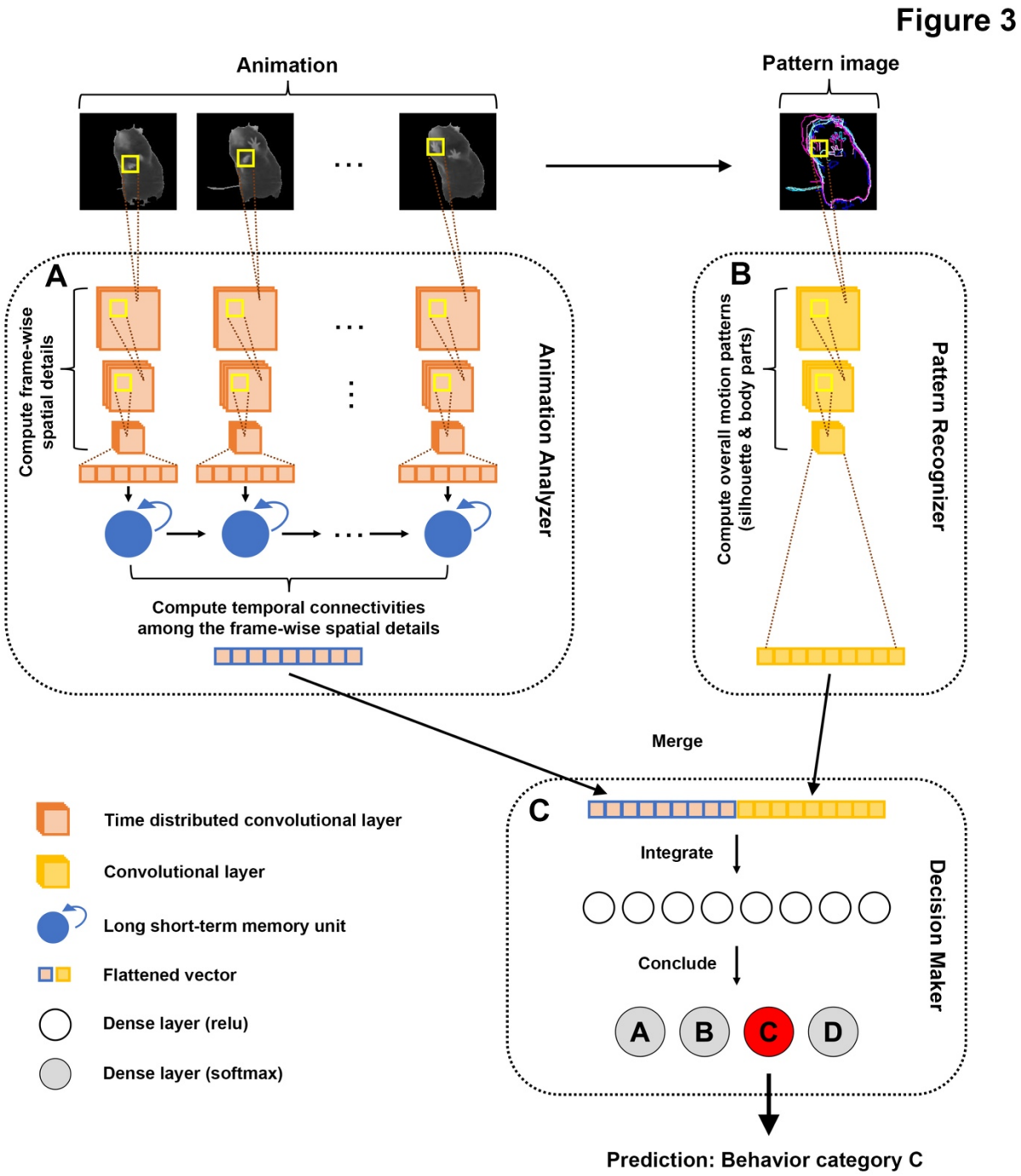


Figure 3. Biological plausible design of the Categorizer.

A. The Animation Analyzer first uses time-distributed convolutional layers to compute all frame-wise spatial details of an animation, and then uses recurrent layers (long short-term memory, LSTM) to compute the temporal connectivity among the frame-wise spatial details.

B. The Pattern Recognizer uses convolutional layers to analyze the pattern image, which is superimposed contours of animal body (and body parts) along the temporal sequence (indicated by gradually changed color) during a behavior.

C. The Decision Maker uses concatenating layers to integrate the outputs from both the Animation Analyzer and the Pattern Recognizer, and then passes the integrated information to dense layers for concluding the behavioral categories.

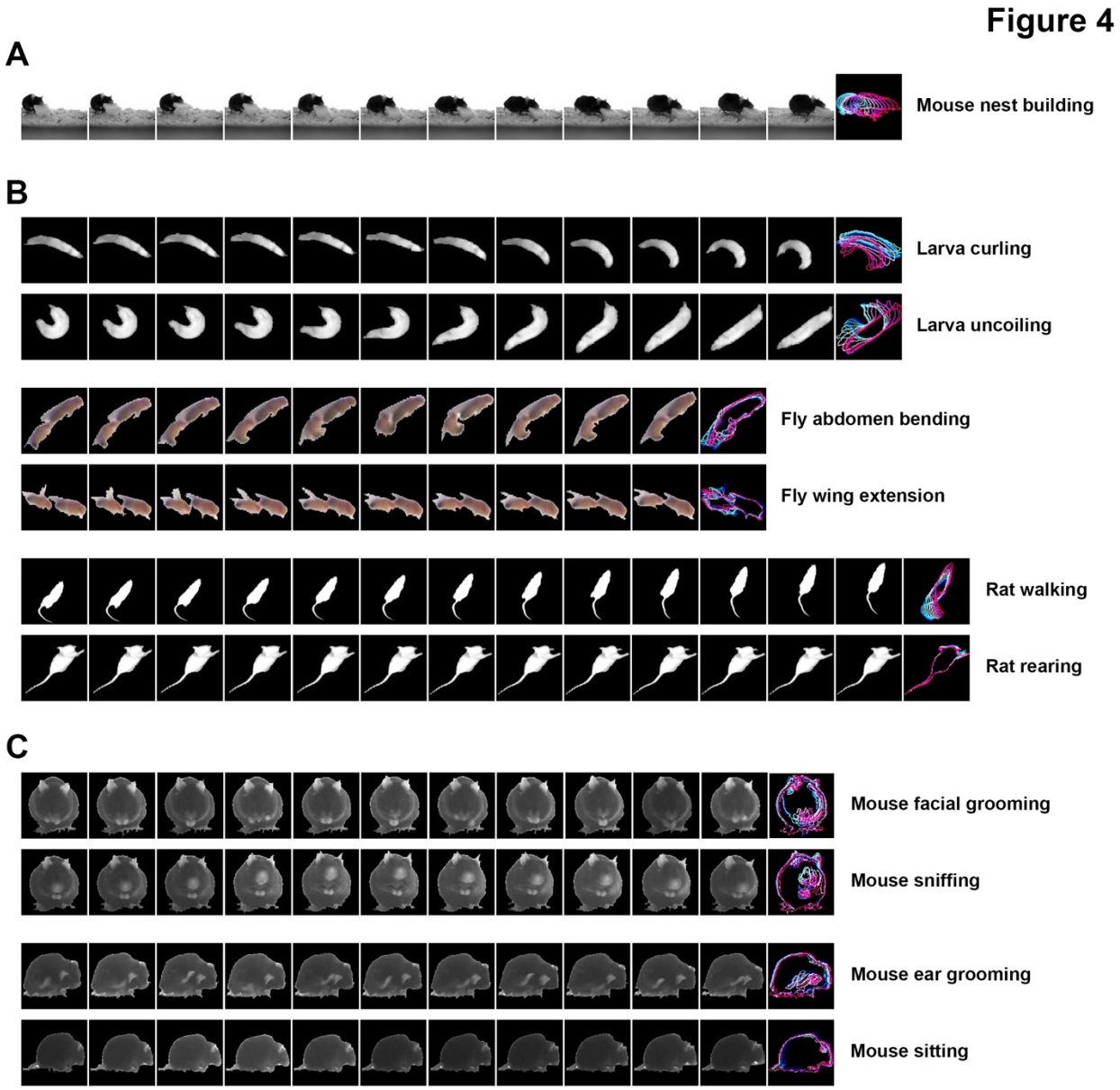


Figure 4. *LabGym* generates stand-alone, visualizable behavioral datasets as benchmarks.

A. Example frames of an animation (background is included) and its paired pattern image, showing a mouse nest building behavior.

B. Example frames of animations and their paired pattern images (body parts are not included), showing (from top to bottom) larva curling and uncoiling, fly abdomen bending (copulation attempt) and wing extension (courtship song), and rat walking and rearing.

C. Example frames of animations and their paired pattern images (body parts are included), showing (from top to bottom) mouse facial grooming, sniffing, ear grooming, and sitting.

Figure 5

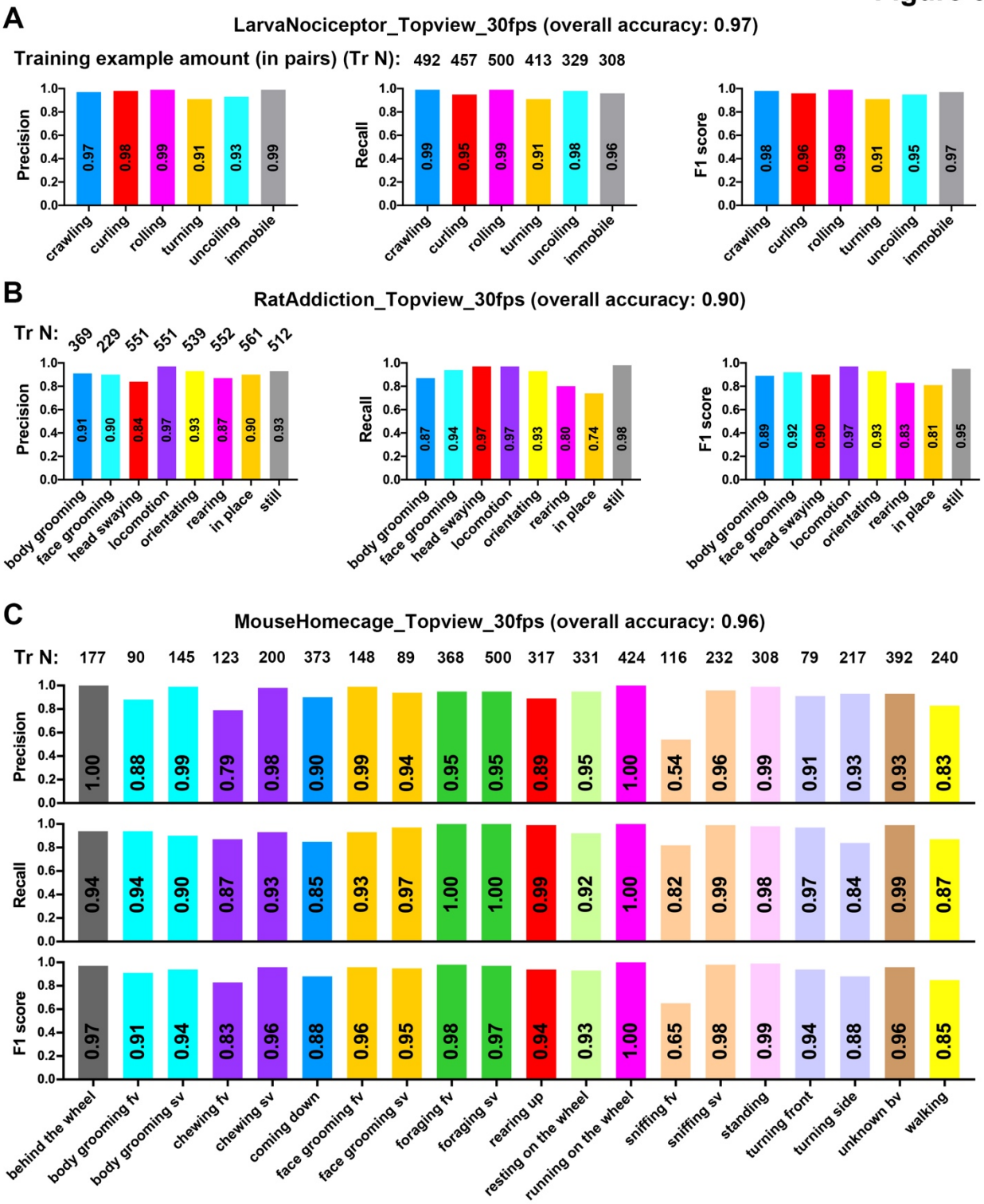
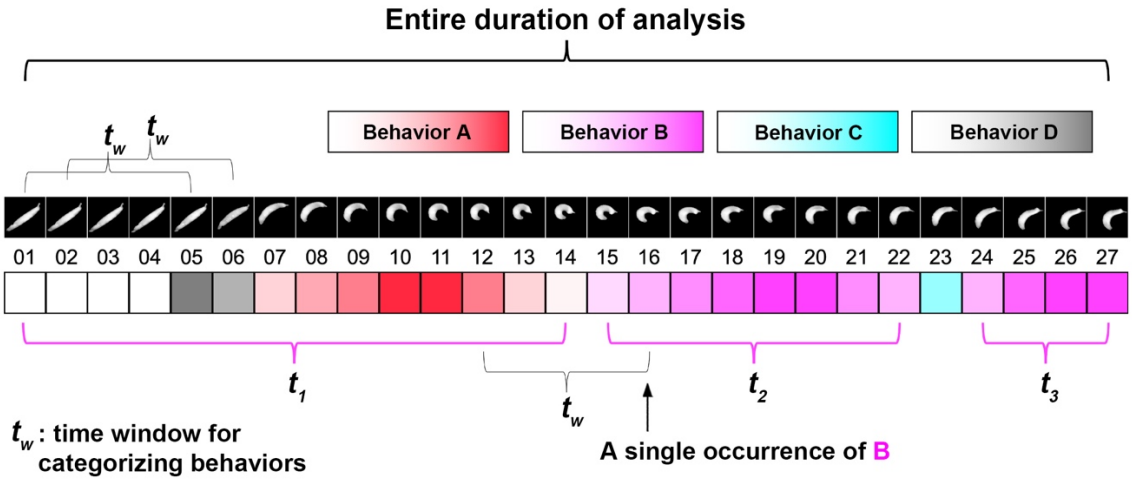
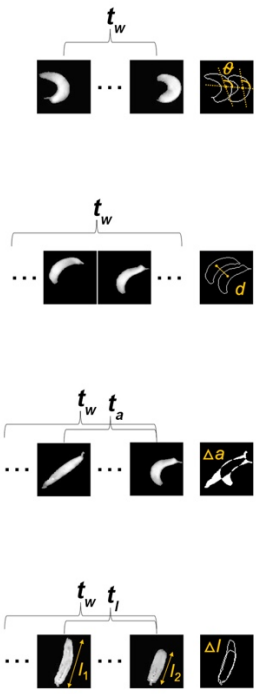

A
LarvaNociceptor_Topview_30fps (overall accuracy: 0.97)
Training example amount (in pairs) (Tr N): 492 457 500 413 329 308
Precision
0.97 0.98 0.99 0.91 0.93 0.99
Recall
0.99 0.95 0.99 0.91 0.98 0.96
F1 score
0.98 0.96 0.99 0.91 0.95 0.97
crawling curling rolling turning uncoiling immobile
B
RatAddiction_Topview_30fps (overall accuracy: 0.90)
Tr N: 369 229 551 551 539 552 561 512
Precision
0.91 0.90 0.84 0.97 0.93 0.87 0.90 0.93
Recall
0.87 0.94 0.97 0.97 0.93 0.80 0.74 0.98
F1 score
0.89 0.92 0.90 0.97 0.93 0.83 0.81 0.95
body grooming face grooming head swaying locomotion orientating rearing in place still
C
MouseHomecage_Topview_30fps (overall accuracy: 0.96)
Tr N: 177 90 145 123 200 373 148 89 368 500 317 331 424 116 232 308 79 217 392 240
Precision
1.00 0.88 0.99 0.79 0.98 0.90 0.99 0.94 0.95 0.95 0.89 0.95 1.00 0.54 0.96 0.99 0.91 0.93 0.93 0.83
Recall
0.94 0.94 0.90 0.87 0.93 0.85 0.93 0.97 1.00 1.00 0.99 0.92 1.00 0.82 0.99 0.98 0.97 0.84 0.99 0.87
F1 score
0.97 0.91 0.94 0.83 0.96 0.88 0.96 0.95 0.98 0.97 0.94 0.93 1.00 0.65 0.98 0.99 0.94 0.88 0.96 0.85
behind the wheel body grooming fv body grooming sv chewing fv chewing sv coming down face grooming fv face grooming sv foraging fv foraging sv rearing up resting on the wheel running on the wheel sniffing fv sniffing sv standing turning front turning side unknown bv walking

Figure 5. Accurate frame-wise categorizations on various user-defined behaviors by different Categorizers.

A-**C**. Bars showing the in-practice test metrics (overall accuracy, precisions, recalls and f1 scores) of *LarvaN* (**A**), *RatA* (**B**) and *MouseH* (**C**). 7 videos for larvae (10 larvae per video), 10 videos for rats (1 rat per video), and 2 videos for mice (1 mouse per video) that were not used for generating the training datasets were used to generate the testing data (animations and their paired pattern images). 2162 pairs (for larvae), 4828 pairs (rats), and 12653 pairs (mice) of data were randomly selected (data with obvious missing body parts were excluded) and then were sorted by experimenters into different folders under the behavior names (building ground truth testing datasets). The ground truth testing datasets were then used for testing the categorization metrics of the 3 Categorizers. In this way, experimenters were blind to the predictions of the 3 Categorizers when sorting data to build the ground truth testing datasets and thus the tests were unbiased.

Figure 6





Count of B = 2 Latency of B = t_1 Duration of B = $t_2 + t_3$

Angle (single occurrence) = $\bar{\theta}_i$ (i in t_w) Speed = $(\Sigma d_i) / t_w$ (i in t_w)

Velocity (V) = max(dt) / t (t: time for the max displacement dt)

Acceleration = ($Vmax$- $Vmin$) / t_v (t_v: time for the change of V)

Distance = Σd_i (i in entire duration of analysis)

Intensity (a) (single occurrence) = $(\Sigma(\Delta a / a_i)) / t_w$ (i in t_w)

Intensity (l) (single occurrence) = $(\Sigma(\Delta l / l_i)) / t_w$ (i in t_w)

Magnitude (a) (single occurrence) = max($\Delta a / a_i$) (i in t_w)

Magnitude (l) (single occurrence) = max($\Delta l / l_i$) (i in t_w)

Vigor (a) (single occurrence) = Magnitude (a) / t_a

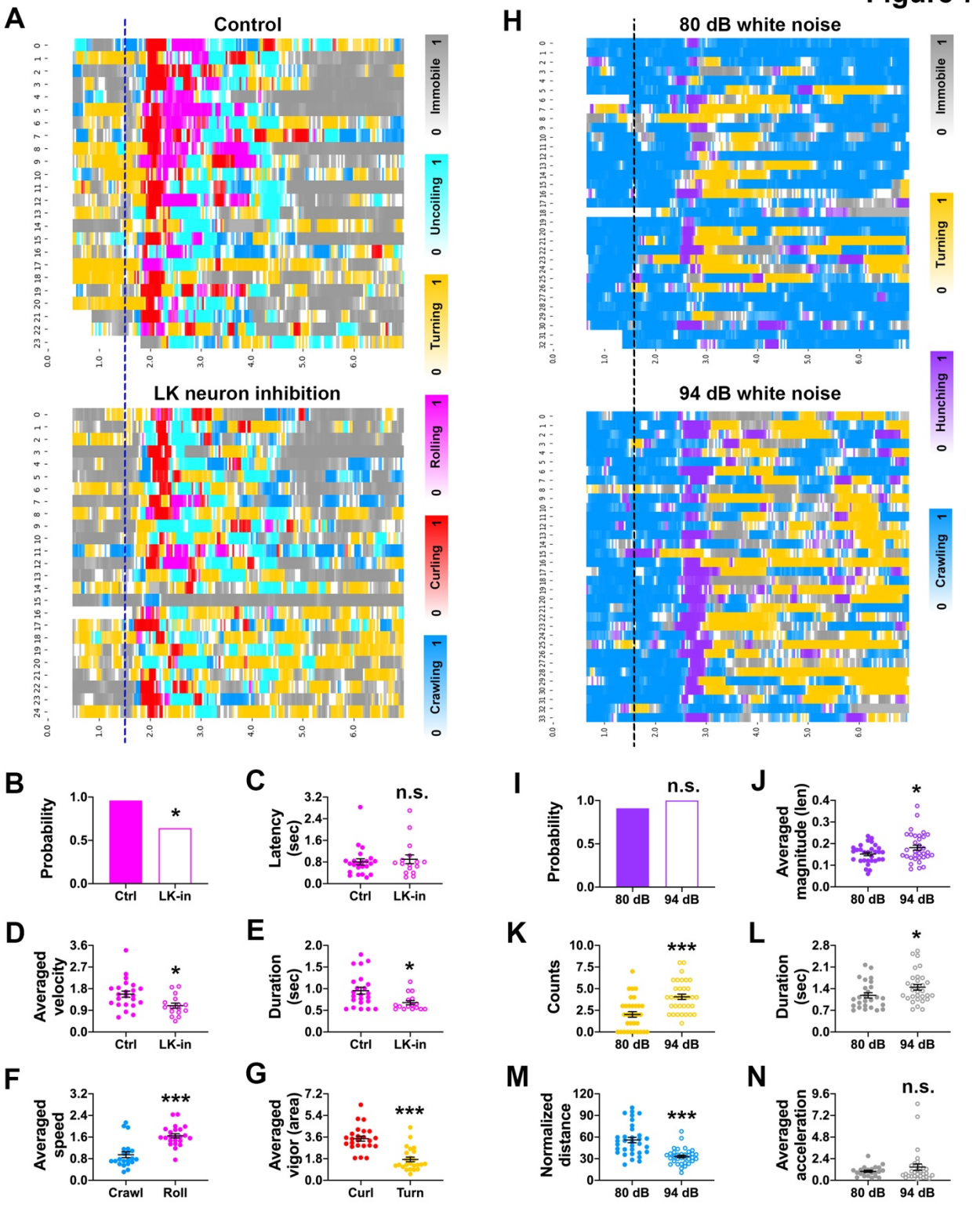
Vigor (l) (single occurrence) = Magnitude (l) / t_l

1

Figure 6. The Quantifier analyzes diverse aspects of behavioral dynamics.

A schematic that shows the process of frame-wise categorizations on behaviors, and how the Quantifier uses the information of behavioral categories and animal foregrounds to calculate 14 behavioral parameters in the current version of *LabGym*.

Figure 7


A
Control
LK neuron inhibition
Immobile
Uncoiling
Turning
Rolling
Curling
Crawling
H
80 dB white noise
94 dB white noise
Immobile
Turning
Hunching
Crawling
B
Probability
Ctrl
LK-in
C
Latency (sec)
n.s.
D
Averaged velocity
E
Duration (sec)
F
Averaged speed
Crawl
Roll
G
Averaged vigor (area)
Curl
Turn
I
Probability
80 dB
94 dB
J
Averaged magnitude (len)
K
Counts
L
Duration (sec)
M
Normalized distance
N
Averaged acceleration

1

Figure 7. The Quantifier captures subtle changes in the behaviors of *Drosophila* larvae.

A and **H**. Raster plots showing the frame-wise categorizations by: **A**. *LarvaN* on control (*w*; *LK*-GAL4 / +; *TrpA1*-QF, QUAS-ChR2^{T159C}, UAS-jRCaMP1b / +) and LK neuron inhibition (*w*; *LK*-GAL4 / +; *TrpA1*-QF, QUAS-ChR2^{T159C}, UAS-GtACR1 / +) groups after nociceptors were optogenetically activated (blue dashed line indicates the stimulation onset); **H**. *LarvaMechanosensor_Topview_30fps* on larvae (*Caton S*) under different intensities of sound stimuli (black dashed line indicates the stimulation onset). Color bars: behavioral categories (color intensities indicate the behavioral probabilities). X-axis represents the time and Y-axis represents the larval IDs.

B-**G**. Parameters calculated by the Quantifier for different behaviors elicited by nociceptor activation. The colors match those in **A**. Data in **F** and **G** are from the control group in **A**.

I-**N**. Parameters calculated by the Quantifier for different behaviors elicited by sound stimulation. The colors match those in **H**.

Probability in **B** and **I** is the fraction of responders (behavioral count > 0) in total animals. n.s., P > 0.05; *, P < 0.05; ***, P < 0.001 (Fisher's exact test for **B** and **I**; unpaired t test for **C**-**G** and **J**-**N**).

Figure 8

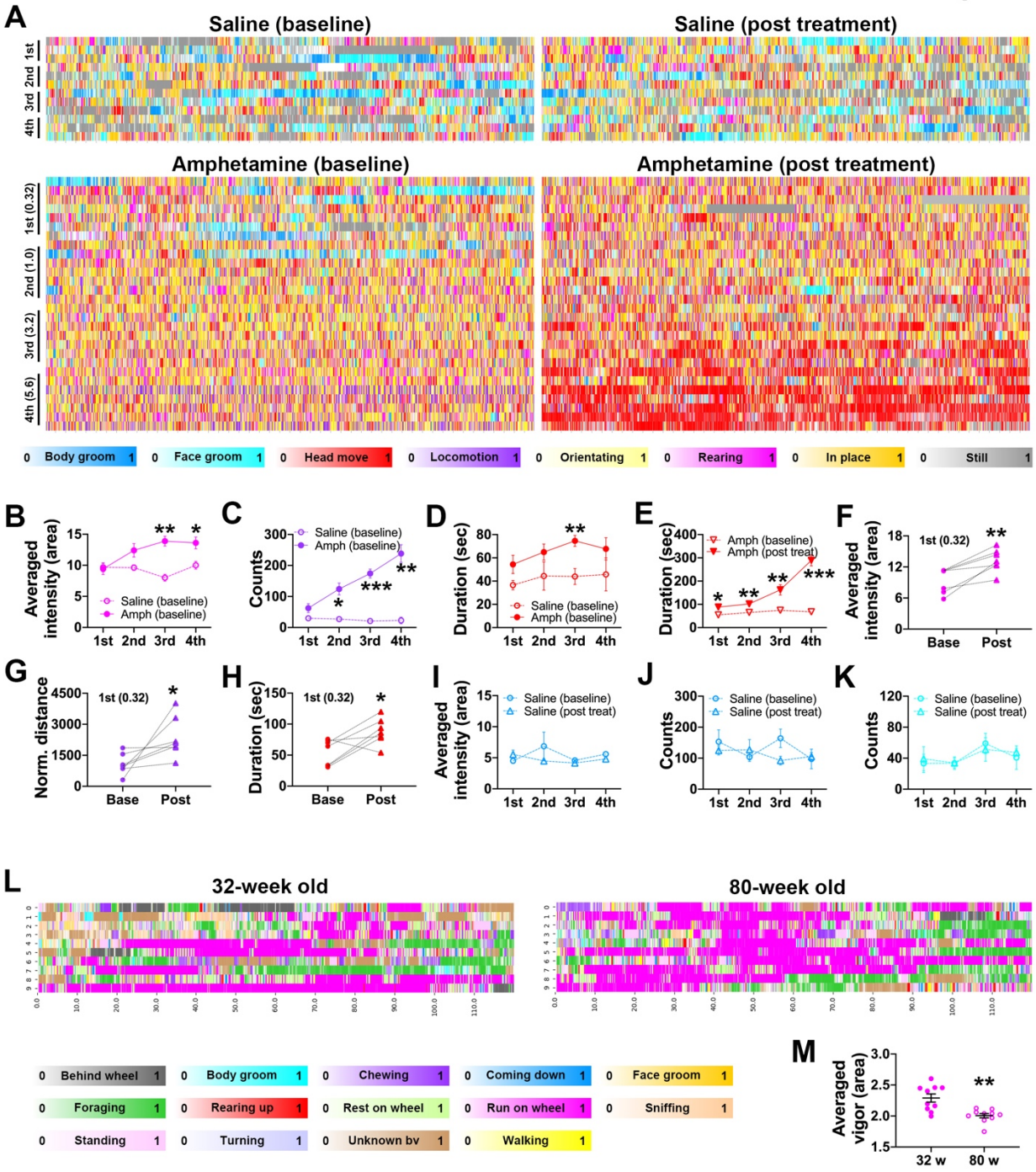


Figure 8. The Quantifier captures qualitative and quantitative changes in rodent behaviors.

A. Raster plots showing the frame-wise categorizations by *RatA* of behavior after saline or amphetamine injection. Plots at the left (baseline) show responses to the first exposure to amphetamine across increasing doses; right (post treatment): repeated injections during a 21-

day sensitization phase and then tested with the 4-dose injections. The injection interval between 4 doses (0.32, 1.0, 3.2, and 5.6 mg / kg for amphetamine) is 30 minutes. Color bars: behavioral categories (intensities indicate the behavioral probabilities). X-axis: time (each tick is 10-sec); Y-axis: each row represents an individual rat.

B-**D**. The dose response curves of rearing intensity (**B**), locomotion counts (**C**) and head movement duration (**D**) in initial drug exposure (baseline) in **A**. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$ (unpaired t test between saline and amphetamine groups).

E. The dose response curves of head movement duration in initial exposure (baseline) and repeated exposure (post treatment, i.e., sensitization) of amphetamine or saline. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$ (unpaired t test between baseline and post treatment groups).

F-**H**. the rearing intensity (**F**), locomotion distance (**G**) and the head movement duration (**H**) in initial (baseline) / repeated exposure (post treatment) group to the same amphetamine dose (0.32 mg / kg). *, $P < 0.05$ (paired t test).

I-**K**. The dose response curves of intensity and counts of body grooming (**I** and **J**) and the duration of face grooming (**K**) in initial (baseline) / repeated exposure (post treatment) group show no difference.

L. Raster plots showing the frame-wise categorizations by *MouseH* for a younger (32-week) or an older (80-week) mouse. Color bars: behavioral categories (intensities indicate the behavioral probabilities). X-axis: time (in second); Y-axis: different random 2-minute time window. Multiple (>100) 2-minute time windows were randomly selected from a 4-hour recording for each mouse, and those containing wheel-running behavior (targeted time window) were sorted out. 10 targeted time windows for each mouse were randomly selected for analysis.

M. The running vigor of the two mice in **L**. **, $P < 0.01$ (unpaired t test).